IRI Total Data Management | CELEBRATING 40 YEARS 2018 The CoSort Company

The Case for DATA MASKING & IRI Capabilities White Paper

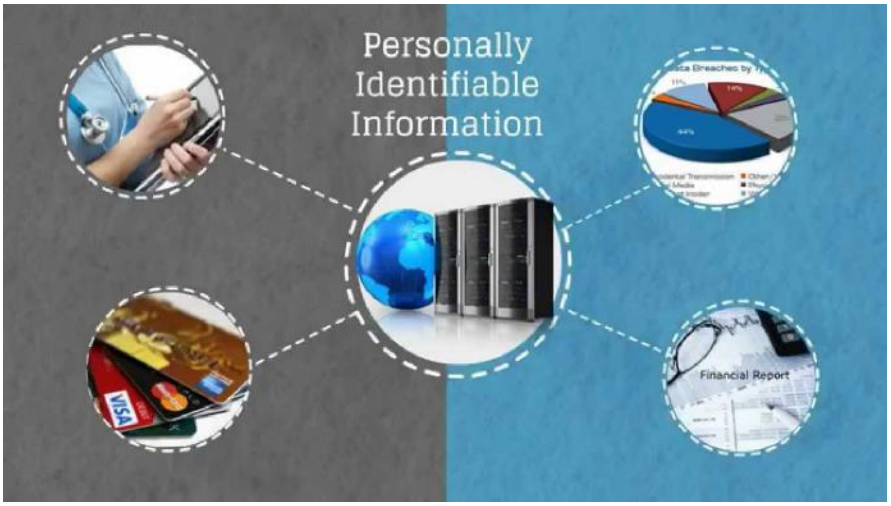
Updated January 2018

Table of Contents

Introduction

Amid the growing regulatory environment for personally identifiable information (PII) worldwide, multiple technology solutions and compliance services have arisen to address PII protection.

Logical encryption in one form or another is a common approach denominator, but most commercial encryption applications are limited by platform, algorithm, ciphertext appearance, implementation complexity, runtime performance, and high costs. Wholesale encryption of sources and devices removes access to non-sensitive data, and leaves everything vulnerable to a single encryption key breach. Thus, there is the need for targeted, versatile, and efficient technology to identify, protect, and audit the remediation efforts applied to PII in different sources.



IRI addresses privacy protection in a data-centric way by extending the field-level parsing and manipulation technology first developed in its CoSort data manipulation package to a fit-for-purpose product called FieldShield, and now also an end-to-end big data management platform called Voracity.

PII is identified, and then secured through encryption, redaction, and other anonymizing data masking functions determined by business needs and/or privacy laws. Its jobs generate audit logs to help verify compliance with data privacy laws. In the Voracity context, data masking can also be embedded into data integration, migration, replication, reporting, and analytic data preparation.

This white paper briefly highlights the regulatory landscape, makes general recommendations, and presents several techniques to help satisfy compliance objectives through data masking. Descriptions and sample screenshots are also provided.

An appendix to this white paper describing ways in which IRI software helps comply with data privacy laws and achieve specific COBIT objectives is available on request.

The Compliance Landscape

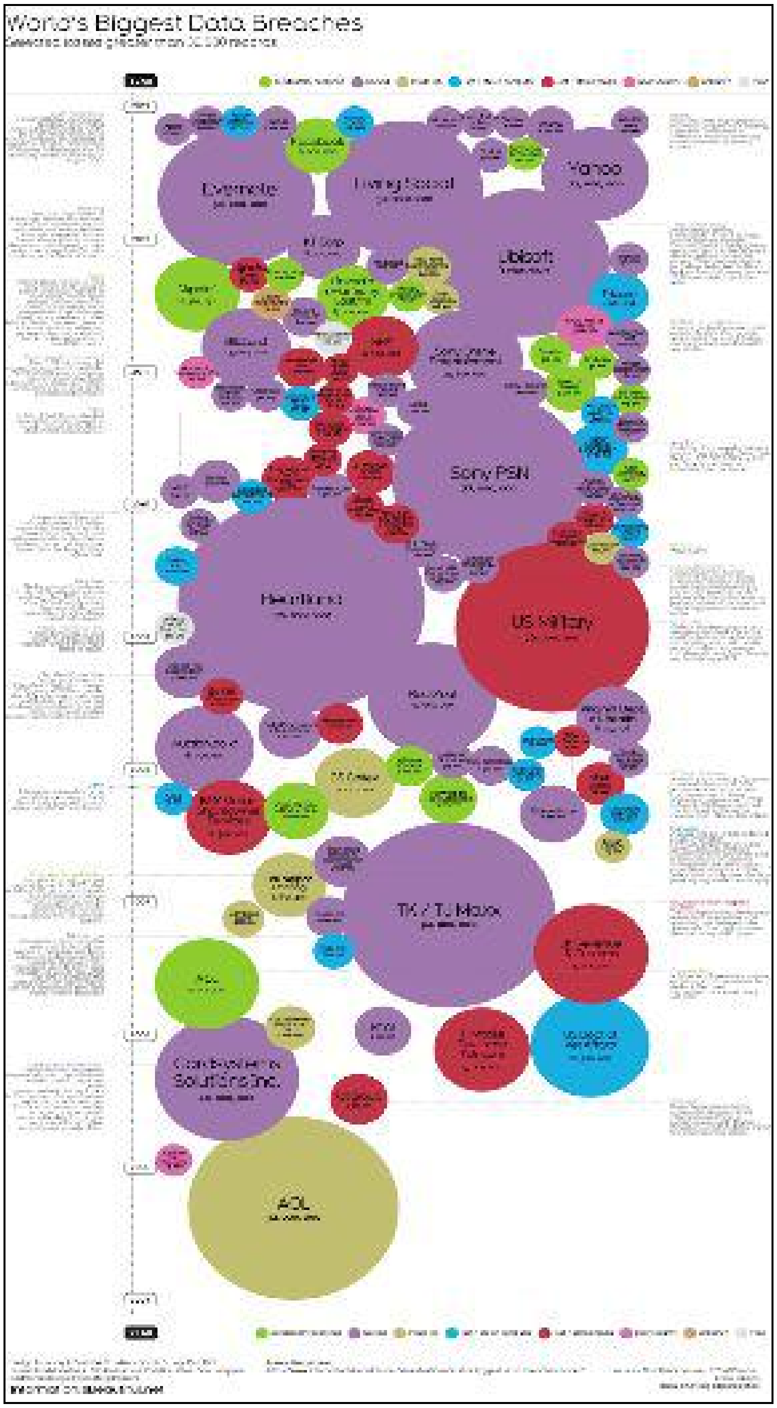
Corporate leaders and information technology executives understand the rewarding possibilities of business intelligence gleaned from their stores of transaction data as well as public sources. Unfortunately, there is a converse, penalty-enforced requirement to protect data from unintended users and uses. Put another way, revenues must also be safeguarded through *safer data*.

A major emphasis is therefore being placed on compliance with government regulations, industry standards, and corporate policies designed to identify and protect data at risk of misuse through disclosure, outsourcing, hacking, theft, and otherwise improper handling.

From an auditing and risk control perspective, executives are scrambling to understand compliance and fund the implementation of controls that will solve problems with data before they surface. From a data perspective, business and IT stakeholders are also discovering that data quality is questionable, and that they often cannot trace data from point A to B without disconnects; i.e., where data starts in the system and ends in reports. Understanding and verifying these data flows are keys to complying with regulations like Sarbanes-Oxley (SOX).

In this compliance era, the common language is now one of risk management. IT departments are now learning that language in many of the same ways business people have always had to – including the hard way. This paper lays out that landscape, and suggests ways to help achieve (and verify) compliance, while still producing rapid, actionable business intelligence.

Let's begin with some of the key considerations and recommendations surrounding risk mitigation in the post-SOX environment:



1) Data = Risk

Data are always at risk -- at risk of hacking, theft, incorrect transmission or modification, being pulled from the wrong sets, and so on. If sensitive data are disclosed, the Privacy Rights Clearinghouse (PRC) posts the incident for the world to forever know what companies allowed its customers' personal information to be compromised. In addition to the embarrassing secrets revealed, lawsuit damages, government fines, and/or data remediation costs can cripple the entire company.

According to the PRC and its Chronology of Data Breaches, roughly one billion personal records have been (reported) compromised in the United States alone since 2005. Much of the PII obtained by identity thieves include Social Security, driver's license, and various account numbers. The PRC's enormous and still-growing list is also comprised of breaches that do NOT expose such sensitive information in order to emphasize the wide array and prevalence of data breaches.

2) Risk Must Be Managed

Managers must understand the risks data represent and then formulate, execute, monitor, and verify plans to remove these risks through reasonable means. SOX adopters are specifically obligated to conduct a formal risk assessment and choose a risk management strategy. Assessing risk is about identifying the strongest intersections between the likelihood of data compromise and its potential impact.

3) Managing Risk with Formal Controls

IT groups are starting to understand the language of "Controls," which are manual or automated steps and software to prevent, detect, and/or correct a problem. Preventive controls are the best (like a door lock). Some tools can do more than one thing; e.g., a firewall prevents and detects. IRI software prevents and corrects detected problems with data through data discovery, masking and auditing. Automating these types of controls is preferable because people make (inconsistent) mistakes.

4) Executives Are Responsible for the Controls

Because information in financial reports (that drove the WorldCom and Enron scandals, and thus, subsequent SOX legislation) can flow through IT department resources, CEOs, CFOs and CISOs need IT's help to attest to the accuracy of the data and the efficacy of the controls. SOX requires this in the financial reporting sphere, but consider also the need to comply with data privacy laws.

GDPR, HIPAA, PCI, and many newer regulations, like the Electronic Privacy Directive, means you must guarantee data could not have been breached, changed, or hacked, and that you must be able to show how your company tries to protect the data. A good control system must also have logging built-in for verifying the use of these controls. Companies who safeguard PII and prove it, not only prevent fines and lawsuits, but enjoy the repeat business of more trusting customers.

5) Deploy Industry-Standard Controls

Using approved protocols for data and access protection supports compliance. Examples include:

- FIPS 140-2 and NSA Suite B encryption techniques

- Committee of Sponsoring Organizations of the Treadway Commission (COSO)
  Principles, reports, and recommendations from a private-sector consortium in the USA founded in 1985 "dedicated to improving the quality of financial reporting through business ethics, effective internal controls, and corporate governance."

- Control Objectives for Information and Related Technologies (COBIT)
  "IT governance framework and supporting toolset that allows managers to bridge the gap between control requirements, technical issues and business risks."

- International Standards Organization (ISO) 17799
  10 controls first published in 2000 comprising the best practices in securing information assets.

- Information Technology Infrastructure Library (ITIL)
  Published by the Office of Government Commerce in Great Britain, ITIL focuses on IT services and is often used to complement the COBIT framework.

- SANS Institute / Center for Internet Security - Critical Security Controls
  20 highest-priority recommendations for cyber defense updated since 2008 by an international, grass-roots consortium of corporate, government and institutional cyber analysts, hackers, solution providers, users, consultants, policy-makers, executives, academia, auditors, etc.

6) Implement Data Governance (People and Rules)

Applications are easy to map to corporate organizational charts when a specific person or group is responsible for it. Data, on the other hand, flow throughout the organization and do not map to specific points on an organization chart. Only now are companies beginning to understand that they cannot map data accountabilities to specific job functions or departments.

The Data Governance Institute asserts that responsibility must be mapped across the organization. This is the goal of those in the "Data Governance Office." Data Governance sets the rules of engagement for people responsible for specifying, designing, implementing, monitoring, testing, and retiring the controls on data. The controls have a life cycle, just like applications and data do.

Data Governance officials are responsible for implementing these rules throughout the life cycle, and corporate IT staff provides input into, and is accountable to, these officials and the agreed-upon rules. It is up to data governance efforts to identify the location and nature (e.g., risk level, and need-to-know classifications) of data at risk.

7) Master Data Management (MDM)

Master data is a unique, core set of transactional data elements (fields) used in many information systems' processes and transactions; i.e. control tables, or reference data in a lookup table. It exists because companies need to share data across departments and business functions, or, in the case of a merger or joint venture, two or more companies must share data across different platforms.

Master data is used for modeling and data quality rules in data migration, cleansing, and stewardship (procedures to prevent and protect disclosure). When a piece of master data that affects many applications and transactions is missing, wrong, inconsistent, or mislabeled, there will be adverse effects downstream. Therefore, good master data management (MDM) is required to keep this data clean, and to standardize master data models amid the relational taxonomy of facet data.

> *Ventana Research defines master data management as the practices and technologies allowing business and IT to define enterprise-wide master or reference data that is linked to the business. According to Ventana Global Research Director David Waddington, master data management enables companies to continue leveraging their current BI, ERP and data warehousing investments.*

Jane Griffin, a Deloitte Consulting partner specializing in data warehousing and business intelligence systems, posits in her article Information Strategy: Building a Data Management Strategy that there are four processes essential to a good MDM strategy:

1) Data Migration and Integration
2) Data Maintenance
3) Data Quality Assurance *and Control*
4) Data Archiving

Master field data are typically stored in a centralized repository (often in support of a Service-Oriented Architecture) for cleansing and sharing -- suggesting a higher stakes need for field-level protection(s). MDM is thus also a data governance issue, and there is a need for it to be protected by the same logic and methods as transactional data.

General Recommendations

1) Locate & Classify Data at Risk

In order to mask PII, you must first know where it is. According to Gartner, only about half the companies needing to mask PII know precisely where these fields or columns are. Data searching, profiling, and classification tools can locate and group PII according to defined traits (including pattern, lookup table, and fuzzy matching attributes) so that data masking rules can be consistently applied. Consistent like-column function application also preserves referential integrity in masked data sets.

2) Structure & Standardize the Data

Flat records (including CSV) and other interchange formats (including XML and JSON) can be a good common denominator for an organization's sources of truth. Mainframe data, and data on the way to and from databases and other applications (e.g., spreadsheets, reporting tools, and web logs) often reside in flat files. Many of them contain PII which can be quickly and differentially masked, and then re-targeted into DB silos, reports, apps, etc.

Data stored in flat files are also easier to define, compress, store, process, report from, and protect than data in tables or web service applications. On secure servers, the files can be rapidly masked on a standalone basis, or in the course of data warehouse extract, transform, and load (ETL) jobs. Big data processing tools like the SortCL program in IRI CoSort run in the file system (thus bypassing the overhead of slower DBMS I/O and SQL encryption functions) and perform multiple functions at once.

If the data are not in structured sources, consider flattening it into structure flat files. The free dark data discovery tool in IRI Workbench can pattern-search, extract, and produce delimited files which can then be masked and re-targeted. Alternatively, IRI also represents technology that can mask the discovered data in unstructured text, MS office, pdf and image files *in-situ*.

3) Embed Protection Operations

There are many products designed to hide, obfuscate, and protect data, including file and disk encryption hardware and software. And, there are many ways to use and prepare data for analytics in integration environments like the data lake, corporate information factory (CIF), operational data store (ODS), and data warehouse (ETL) environments. But the two technologies rarely meet. Yet when:

- data volumes are growing
- production windows are shrinking
- resources are finite

data transformation and protection should be combined in the same program and I/O pass to save time and money. This can be done in the IRI CoSort product or IRI Voracity platform.

4) Protect Master Data

Risk is centralized in a single place when users move master data to a central repository where it is easier to clean and supports SOA. Consider also the need-to-know and encryption requirements in the case of outsourcing and testing when the master data are “all in one basket.” Special security filters may thus be needed when storing and moving those values, too.

FieldShield, CoSort and Voracity -- as well as policy-based dynamic DB masking in IRI Chakra Max -- can enforce need-to-know rules on master data through user-specific security functions. The CoSort Sort Control Language (SortCL) program behind FieldShield, CoSort and Voracity standardize, or “master” field names and attributes, as well as identify anomalies, transform, filter, cleanse, and report against that data.

5) Use Safe Test Data

Test data has many uses, including:

- Application Development
- Range Testing
- Stress Testing
- Format Sharing (Outsourcing)
- Database Simulation / Population
- Benchmarking

Common techniques for the generation of useful test data have included custom programs, shareware tools for column generation, and taking selected snippets of production data. This last method provides the true essence of the field elements and the actual look and feel of the file or table containing them. However, production data is an inherently unsafe, if not prohibited, source of test data.

While IRI FieldShield and CoSort software can morph production data to make it safe for disclosure and useful for testing, sometimes that production data does not yet exist or cannot be used. In such cases, a test data generation tool is needed. IRI RowGen software uses the same data definition syntax to randomly generate or select data from only the metadata of production tables or files, accurately simulating their value ranges, target structures, and primary-foreign key relationships.

6) Verify Compliance

Keep performance and transaction records of data processing, reporting, protection and generation steps. Statistics should be kept in secured, query ready audit logs that include:

- Runtime information (e.g., when the job was run and how long it took)
- Application statistics (e.g., number of records read, transformed, masked, etc.)
- Source and target names (also resolved from environment variables)
- All field specifications (name, position, size and data type attributes)
- All field and file manipulations (by showing the actual job script used)

All IRI products create these trails to report the actions taken and support compliance verification.

“Safe Field” Techniques

Field-level filtering, transformation, encryption and other data masking techniques such as those available in tools like FieldShield allow data governance officials and IT staff to produce data views or persistent targets that can leave the office without compromising PII. Masking column values is also faster and more useful than hiding whole rows, files, tables, DBs, etc. because the protected data can still be used, and concealed or revealed in very granular ways.

These techniques help, for example: health care companies to achieve compliance with HIPAA regulations by de-identifying or shrouding medical and personal data; to satisfy government agency and contractor needs to encrypt sensitive field values in various ways; and, to allow payment card processors to adhere to PCI standards for the protection of PII (like names, account numbers and expiry dates, home addresses, and telephone, credit card, and social security numbers).

Column & Row Filtering (Deletion)

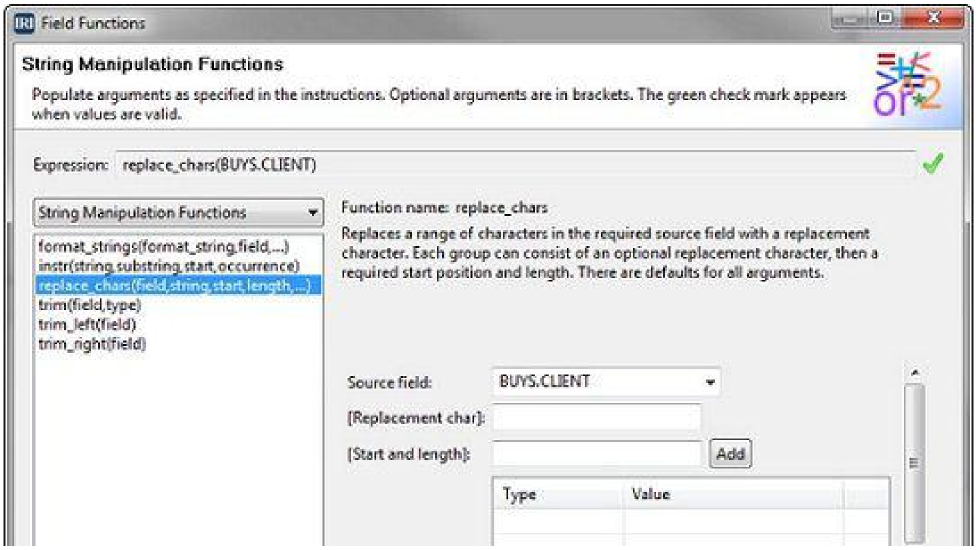
Select which input fields will go to output reports, tables, and hand-off files on a need-to-know basis. Both IRI FieldShield and CoSort achieve this through conditional criteria. This allows you to withhold the display of one or more columns by not specifying its name in the output target(s) pursuant to data privacy laws and/or internal entitlement rules.

Anonymization

Remove the individualizing characteristics of data so that a person or item stored in the original field cannot be identified. Once anonymized, the data cannot be linked to any source. The benefit of anonymization over filtering or (non-format-preserving) encryption is that the original field layout (position, size, and data type) can remain the same, and thus remain realistic for test or demo use.

There are several ways IRI software users can anonymize data, including:

- using mathematical expressions on numeric data
- character shifting or bit manipulation
- data type translation
- redaction (below)


Field Functions

String Manipulation Functions

Populate arguments as specified in the instructions. Optional arguments are in brackets. The green check mark appears when values are valid.

Expression: replace_chars(BUYS.CLIENT) ✓

Function name: replace_chars

Replaces a range of characters in the required source field with a replacement character. Each group can consist of an optional replacement character, then a required start position and length. There are defaults for all arguments.

Source field: BUYS.CLIENT

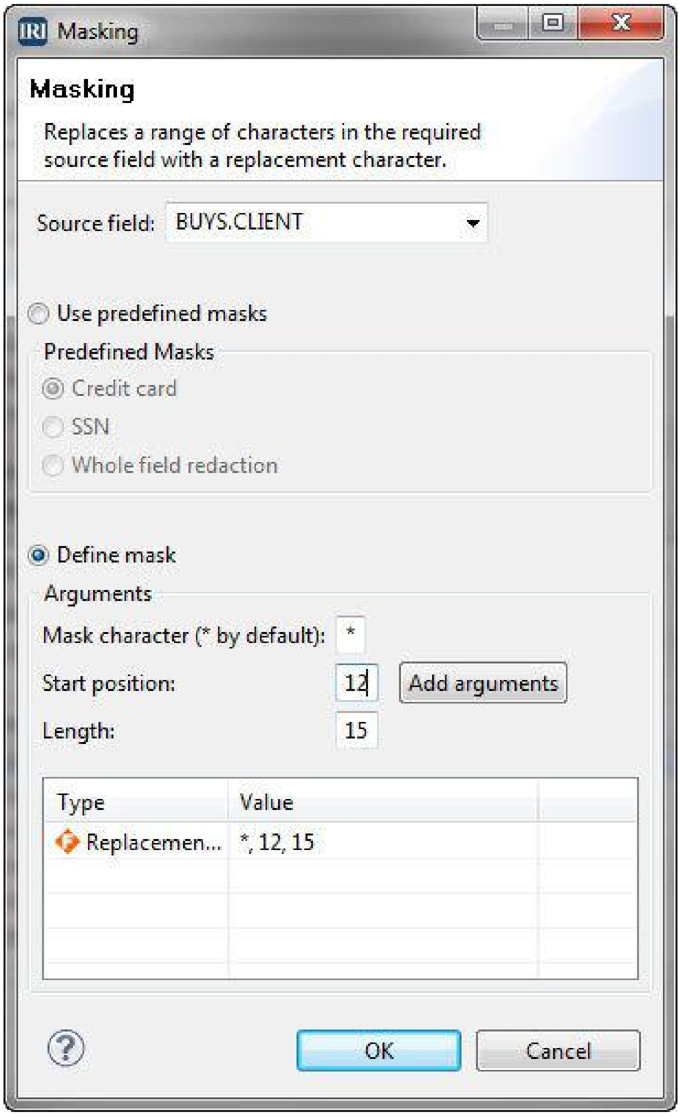
[Replacement char]:

[Start and length]:

String Manipulation Functions
format_strings(format_string,field,...)
instr(string,substring,start,occurrence)
replace_chars(field,string,start,length,...)
trim(field,type)
trim_left(field)
trim_right(field)
Add
Type
Value

Redaction / String Masking (Non-Recoverable)

Data redaction, or character (string) masking, is a form of irreversible anonymization that involves covering one or more bytes of a column (field) value with a chosen obfuscation character, permanently blacking out some or all of the original field value. This can be applied to the field based on conditions so that only those values meeting the defined test are redacted.



Masking

Replaces a range of characters in the required source field with a replacement character.

Source field: BUYS.CLIENT

○ Use predefined masks

Predefined Masks

● Credit card
○ SSN
○ Whole field redaction

● Define mask

Arguments

Mask character (* by default): *

Start position: 12

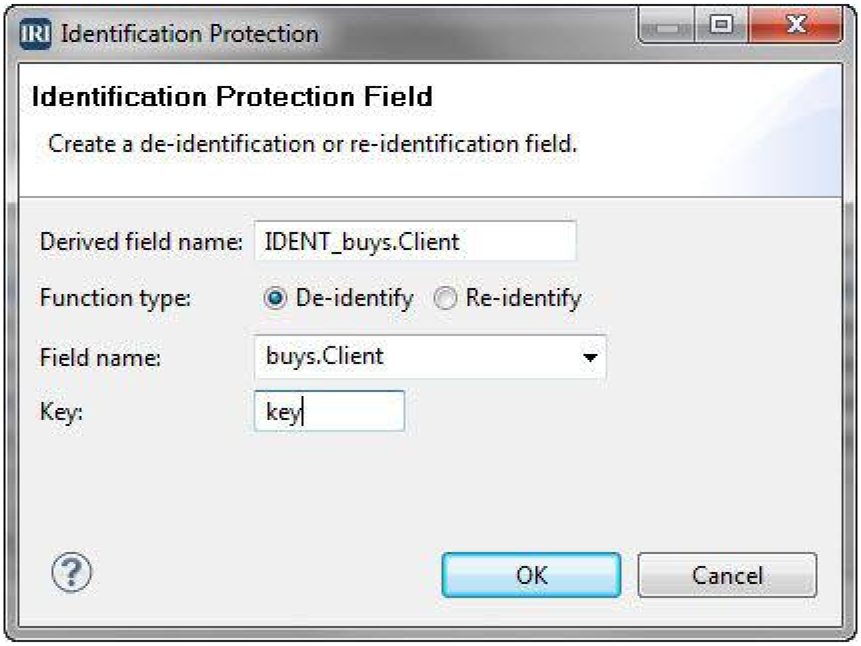
Length: 15

Type	Value
Replacemen...	*, 12, 15

Data masking dialog in the free IRI Workbench (Eclipse IDE) supporting FieldShield, CoSort, Voracity, et al

ASCII De-Identification (and Re-Identification) (Recoverable)

Change the individualizing characteristics of data so that a person or item stored in the original field cannot be identified, but nonetheless remain individualized so that it can be followed (safely) through different departments, and then if necessary, re-identified. This IRI software function works like encryption, but is not as secure algorithmically. Its advantage is therefore relative speed in volume.



Identification Protection

Identification Protection Field

Create a de-identification or re-identification field.

Derived field name: IDENT_buys.Client

Function type: ● De-identify ○ Re-identify

Field name: buys.Client

Key: key

De- and re-identification dialog for ASCII field values in the IRI Workbench GUI for FieldShield, CoSort and Voracity

Encryption and Decryption (Recoverable)

The encryption of data in files, databases, and on disks and other media has been practiced for years. According to expert James C. Foster, author of Look Before Leaping into Database Encryption:

> *Encryption is a powerful security tool, and nearly every compliance standard or industry regulation addresses data security in some manner, often at least implying a role for encryption. For instance, the Gramm-Leach-Bliley Act (GLBA) requires organizations must "insure the security and confidentiality of customer records and information," and California's SB 1386 breach-notification law states that any breach of the security of unencrypted personal information must be disclosed.*
>
> *Here are some simple guidelines that will help you secure your database without impeding the business you're trying to protect:*
>
> - *Never encrypt foreign or super keys (encrypted keys used for indexing could cause usage and performance issues).*
> - *Use symmetric over asymmetric cryptography when available (again, for performance).*
> - *Full database encryption is rarely advised or a feasible option. Security best practices would teach you to encrypt everything with multiple keys and differing algorithms. However, the significant performance hit you must selectively choose.*

- ***Encrypt only sensitive data columns. This is typically all that is required or recommended by regulations and, after all, is what needs protection.***

Determining key fields or data elements is a daunting task and should be driven by compliance and threat mitigation. Since no regulation comes right out and states "X" columns must be encrypted, it falls back on good judgment. Identifying your most sensitive data and all the places it may reside, from primary databases to backups, is one of the toughest parts of implementing encryption, which is why companies may attempt to solve the problem by deciding to simply, if misguidedly, "encrypt everything."

Choosing an appropriate encryption method also depends on your data. If it mainly consists of images and Web content, then a weaker algorithm -- such as DES or SSL -- may be adequate. However, if you are storing personally identifiable customer information or the company design for a nuclear disintegrator, choose strong encryption with a larger key space, such as AES, Blowfish or 3DES.

The choice of encryption products is improving, as database encryption has made significant strides in the past few years, and the market has continued to mature.
…

While compliance pressures have stoked keen interest in database encryption, it doesn't solve all database security concerns, which are at least as much about preventing abuse by privileged insiders as external attackers. There are no shortcuts. Hastily implementing database encryption simply to comply or assuming it alone will make your data secure will cost extra time, money, manpower and brainpower better spent elsewhere.

Encryption is the most secure of the field protection techniques presented. It uses technology that conforms to US government standards (like FIPS) for hiding the contents of sensitive data. On output, the fields are in a largely unusable form until decryption with the proper key occurs.

Field-level encryption is worth considering because of its flexibility and performance benefits:

- Only sensitive fields are encrypted; remaining fields and disk are readable
- Different encryption keys and libraries can be used on different fields
- Common encryption functions and keys supports referential integrity
- Different field protection methods can be used simultaneously
- Encryption can occur during routine data transformation and reporting
- Computing resource overhead is nominal

IRI offers several choices for encrypting fields in multiple data sources, including database columns, flat files, etc. FieldShield, CoSort and Voracity users can encrypt with format-preserving AES-256, AES-128, FIPS-compliant OpenSSL, 3DES, and GPG, or their own runtime-linked functions.

IRI's Suite B-compliant AES implementation can be used with user-provided passphrases or files (encryption keys). The pass phrase acts as a seed to a function that generates a Secure Hash Algorithm hash digest to derive the encryption key.

Database vendor encryption products provide fewer functional choices, can be cumbersome and expensive to implement, and cannot be used on other databases or flat files.

By way of application, consider the Final HIPAA Security Rule enacted in 2003 governing the protection of electronic private health information (EPHI):

> *Section 164.312, **Technical Safeguards**, contains provisions extracted from two sections of the proposed rule: Technical Security Services and Technical Security Mechanisms. **Covered entities must implement:***
>
> > ***Technical policies and procedures for access control on systems that maintain EPHI. These systems must allow for unique user identification** and include an emergency access procedure for obtaining necessary EPHI during an emergency. Addressable specifications include automatic logoff and encryption and decryption, which is defined as **a mechanism to encrypt and decrypt EPHI**.*

With control at the field level, multiple encryption libraries and passphrases can be used for field-specific need-to-know decryption entitlements.

> - *Transmission security, including two addressable specifications:*
>
>   *1. Integrity controls -- security measures to ensure that electronically-transmitted PHI is not improperly modified without detection until disposed of, and*
>
>   *2. Encryption. Designation of encryption as an addressable specification is a key departure from the proposed rule, which explicitly required encryption when using open networks. Covered entities now must determine how to protect EPHI "in a manner commensurate with the associated risk." **Covered entities are encouraged in the Rule's preamble to consider use of encryption technology for transmitting EPHI, particularly over the Internet.** The key reasons cited by HHS for this change are the cost burden for small providers and the current lack of a simple and interoperable solution for email encryption.*

IRI makes low-cost encryption another option, along with filtering, anonymization, and pseudonymization at the field level, while running routine data manipulations and reports.

> - *Hardware, software, and/or procedural methods for providing audit controls.*

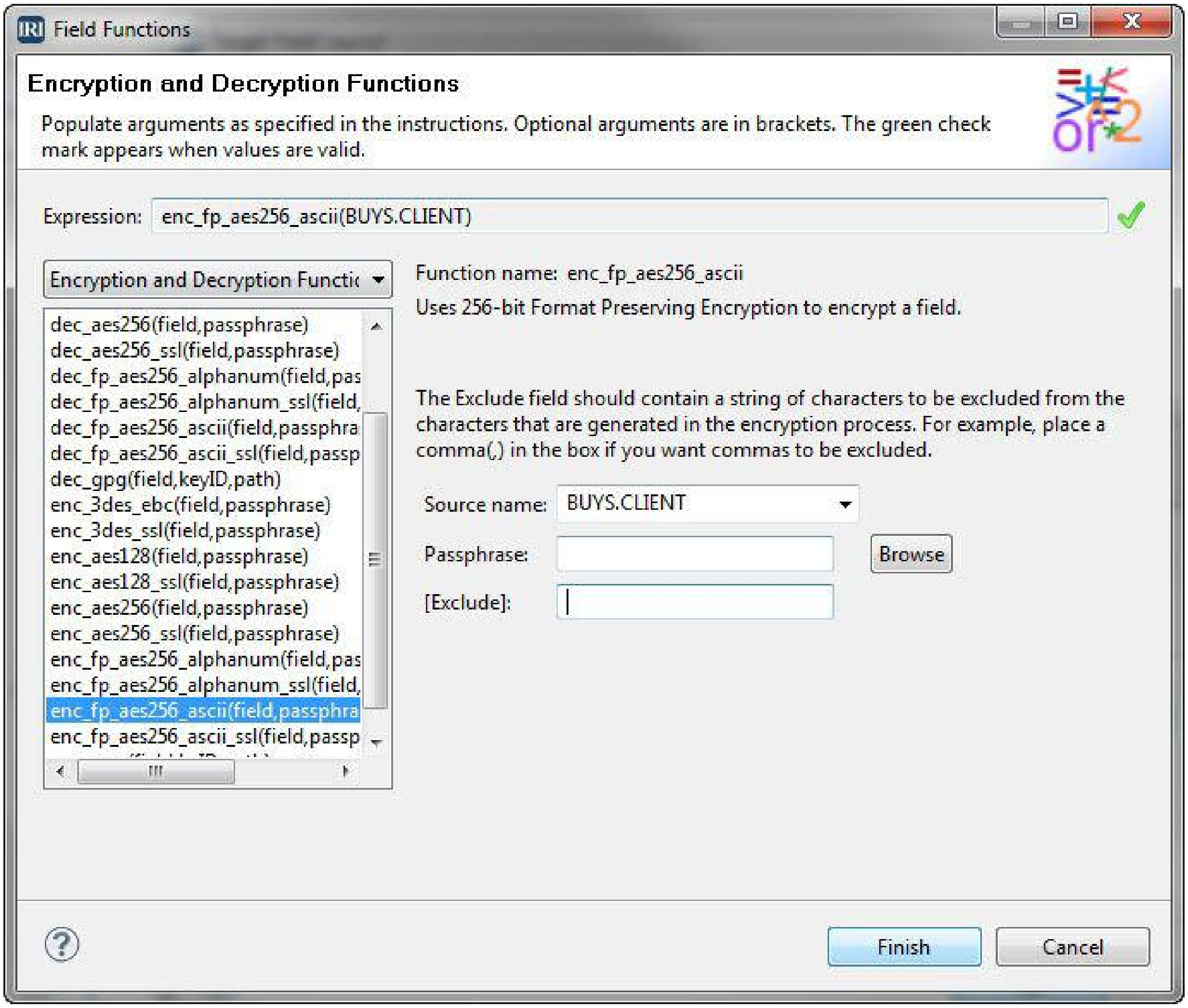
FieldShield and CoSort/Voracity audit records include the full job script, along with the path and name of the encryption libraries. The secure audit log can be used to query and display what, when, how, and by whom the PHI field data was encrypted (and otherwise protected and/or transformed).

> - *Policies and procedures to protect EPHI from improper alteration or destruction to ensure data integrity. This integrity standard is coupled with one addressable implementation specification for a mechanism to corroborate that EPHI has not been altered or destroyed in an unauthorized manner.*

Data that does not decrypt with the proper encryption key suggests that the decrypted field has been compromised. This can be traced in logs that track when the file was processed for field encryption.

- *Person or entity authentication, which requires the covered entity to implement procedures that verify that a person or entity seeking access to EPHI is the one claimed to be doing so.*

Passphrases are used to generate keys for encryption and decryption of field data. Therefore, only the person or entity in possession of the right libraries and passphrase(s) can encrypt or decrypt the field(s). The passphrase can exist inside the job script, implicitly through an environment variable (effectively hiding the pass phrase), or stored in a file within a permissions-restricted directory.



Encryption and Decryption Functions

Populate arguments as specified in the instructions. Optional arguments are in brackets. The green check mark appears when values are valid.

Expression: enc_fp_aes256_ascii(BUYS.CLIENT) ✓

Encryption and Decryption Functic

dec_aes256(field,passphrase)
dec_aes256_ssl(field,passphrase)
dec_fp_aes256_alphanum(field,pas
dec_fp_aes256_alphanum_ssl(field,
dec_fp_aes256_ascii(field,passphra
dec_fp_aes256_ascii_ssl(field,passp
dec_gpg(field,keyID,path)
enc_3des_ebc(field,passphrase)
enc_3des_ssl(field,passphrase)
enc_aes128(field,passphrase)
enc_aes128_ssl(field,passphrase)
enc_aes256(field,passphrase)
enc_aes256_ssl(field,passphrase)
enc_fp_aes256_alphanum(field,pas
enc_fp_aes256_alphanum_ssl(field,
enc_fp_aes256_ascii(field,passphra
enc_fp_aes256_ascii_ssl(field,passp

Function name: enc_fp_aes256_ascii

Uses 256-bit Format Preserving Encryption to encrypt a field.

The Exclude field should contain a string of characters to be excluded from the characters that are generated in the encryption process. For example, place a comma(,) in the box if you want commas to be excluded.

Source name: BUYS.CLIENT

Passphrase:

[Exclude]:

FieldShield encryption and decryption dialog in IRI Workbench

Pseudonymization (Recoverable and Non-Recoverable)

One of the best ways to protect the most identifying piece of personal information – someone's name – is to use a fake name, or pseudonym. IR software provides two methods for replacing the original value of a field with a substitute value:

Non-recoverable pseudonymization

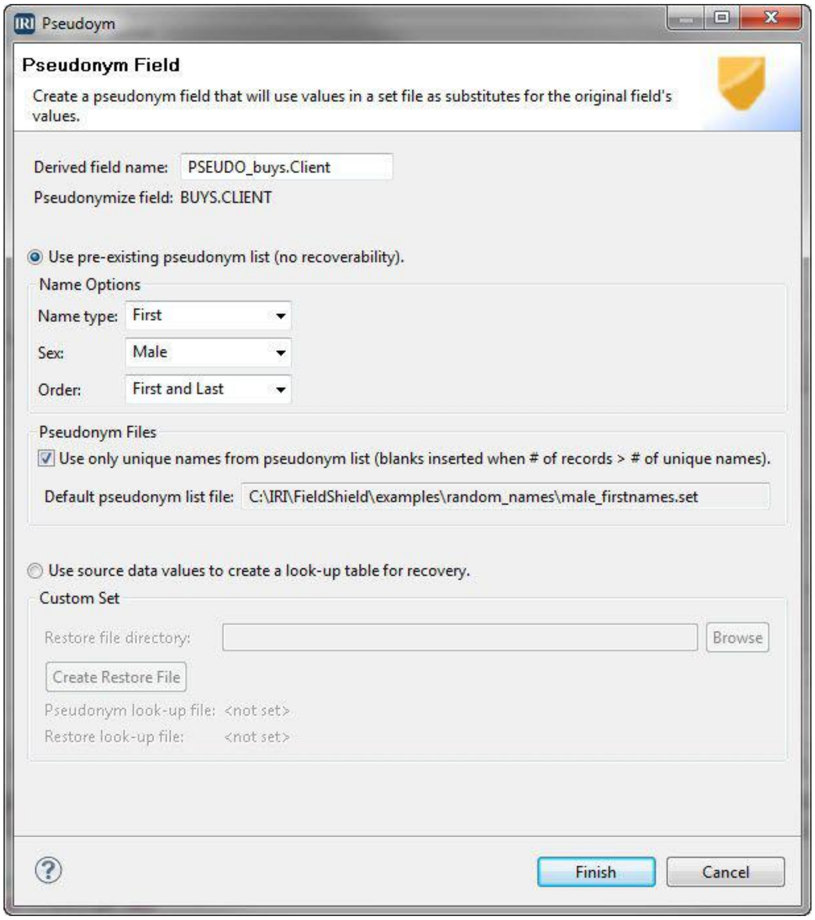
This function performs a random lookup from a supplied list (single-column set) file of names. The user can choose among common western first, last or both names, as well as gender, to use for the name column in the input source. This allows the names to appear real without compromising anyone's identity. Outside the IRI Workbench GUI, users can specify other set files to replace their original values with values of their choosing. This approach does not allow the original names to be restored because there is no matching value association.

Recoverable pseudonymization

This function builds a sorted lookup (two-column set) file containing the original names from the input source, and substitution values randomly selected from that data to be paired with the original values.

This effectively creates a shuffled list of real names so that there is no direct association between the original people and their attributes within other columns. A second, reversed list is created so that recovery is possible.

For additional security, other protection functions can be applied to other source columns. This approach supports greater realism and the ability to restore real names in their original rows.



Pseudonym Field

Create a pseudonym field that will use values in a set file as substitutes for the original field's values.

Derived field name: PSEUDO_buys.Client

Pseudonymize field: BUYS.CLIENT

● Use pre-existing pseudonym list (no recoverability).

Name Options

Name type: First

Sex: Male

Order: First and Last

Pseudonym Files

☑ Use only unique names from pseudonym list (blanks inserted when # of records > # of unique names).

Default pseudonym list file: C:\IRI\FieldShield\examples\random_names\male_firstnames.set

○ Use source data values to create a look-up table for recovery.

Custom Set

Restore file directory:

Pseudonym look-up file: <not set>

Restore look-up file: <not set>

Pseudonym value creation and specification dialog in IRI Workbench

Randomization (Non-Recoverable)

Replacing real values with random values is yet another approach to shielding PII from disclosure while maintaining the original structures of the input sources. This process is not reversible and thus cannot be used to maintain referential integrity. IRI provides two methods for replacing the original value of a field with a substitute value:

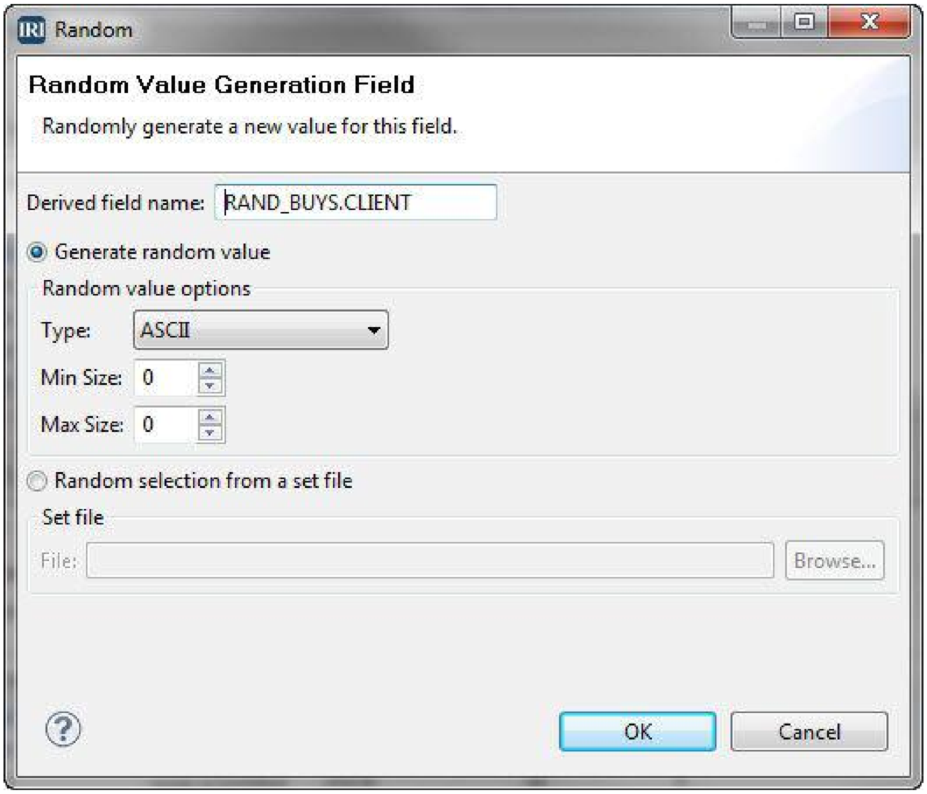
Random value generation

Source field values are replaced with randomly generated values of a given data type. Using random data, especially numbers, can make the protected field look real, while not requiring the use of any real data or protective overhead.

Random value selection

Source field values are replaced with randomly selected values in a lookup ("set") file containing:

- an alphanumeric list of values: the values appear in output exactly as they appear in the set file, producing realistic-looking data and the values that appear in output contain only the numbers, or a specified range of numbers, in the set file
- date values and ranges: only listed dates, or valid dates within a specific range, appear



Random

Random Value Generation Field

Randomly generate a new value for this field.

Derived field name: RAND_BUYS.CLIENT

● Generate random value

Random value options

Type: ASCII

Min Size: 0

Max Size: 0

○ Random selection from a set file

Set file

File:

Random value creation in IRI Workbench

General Field Protection Tool Notes

All of the above techniques – from field removal to randomization – can operate independently, on one or more tables and files at a time in IRI FieldShield, CoSort or Voracity. The same protection functions can also run with or without data transformation or reporting activities in the same job script and I/O pass. With any IRI product, there is no minimum or maximum restriction on the size or number of data sources to be protected.

Conclusion

Your organization manages personally identifiable information (PII). Your data governance efforts must prevent the kinds of data disasters posted at the Privacy Rights Clearinghouse. You must comply with industry and government data privacy rules.

Since you cannot eliminate PII, you have to discover it, protect it, and verify that you protected it. Then you have to continue monitoring and addressing data risks going forward.

At protection time, technology choices are difficult. Traditional encryption of entire databases, files, disks, or devices is inefficient (especially in volume), restricts access to non-sensitive data, and is subject to complete exposure from a single password breach. Many data masking methods are insecure, complex, expensive, or render the protected data unusable for testing.

Moreover, with current methods you may not get:

- an audit trail detailing how you managed risk -- forcing a costly validation exercise
- a separation of encryption and key management (should either be compromised)
- the ability to simultaneously apply multiple protections to multiple data sources
- the ability to combine data protection with other data processing operations

“Solutions that provide encryption at the file, database field, and application level provide the highest level of security while allowing authorized individuals ready access to the information. Decentralized encryption and decryption provide higher performance and require less network bandwidth, increase availability by eliminating points of failure, and ensure superior protection by moving data around more frequently but securely.”

-- Gary Palgon, Enterprise Systems Journal

IRI Technologies

Individual software products in the IRI Data Protector suite, including

IRI FieldShield	finds, masks and audits PII in structured file, RDB/NoSQL, or HDFS
IRI CellShield	finds, reports on, masks, and audits changes to PII in Excel sheets
IRI DMaaS	data masking as a service, including PII identification and post-fix audits
IRI RowGen	puts realistic and referentially correct test data in DBs, files and reports
IRI Chakra Max	monitors, blocks, masks, and audits (firewalls) data in high traffic DBs
IRI Workbench	free Eclipse GUI for data profiling, DB ops, and IRI job management
New: Unstructured Masking	in-situ and on-the-fly redaction and pseudonymization in .pdf, MS Office, image files, JSON (with floating PII), etc.

profile and protect sensitive data, and facilitate privacy law compliance with a broad array of static (SDM) or dynamic data masking (DDM) functions for PII in databases and files, structured and not.

FieldShield, for example, marries a familiar Eclipse GUI with a cross-platform 4GL and executable, to:

- Search, classify, and model PII across a wide range of data sources
- Encrypt and decrypt with multiple built-in (or your own) libraries
- Pseudonymize, encode, hash, randomize, and tokenize
- Filter or redact fields or records based on conditions

FieldShield produces query-ready XML audit logs to document and verify every protection. It can also mask DB subsets for testing. However, consider IRI RowGen for generating safe, referentially correct test data from scratch instead, especially if you cannot access production data or need better data.

FieldShield is also an included component of the IRI Voracity platform, which consolidates enterprise data lifecycle management activities: data discovery, integration, migration, governance, and analytics. FieldShield also shares the same metadata and Eclipse GUI with other IRI tools, including IRI CoSort (for data transformation and reporting), IRI NextForm (or data and database migration and replication), and IRI RowGen (for test data generation from scratch). Their common data definition and manipulation metadata allow you to mask data as part of many enterprise information management (EIM) operations.

In summary, FieldShield -- or optional data masking services from IRI or experts you choose -- secures sensitive data inside and outside the firewall. The result is compliance with data privacy laws, and also:

- nullification of future data breaches (by virtue of having data pre-masked)
- support for the risk and controls framework of your enterprise; e.g., DLP and stewardship goals
- greater consumer confidence in your organization’s ability to protect PII
- seamless interoperation with other data management activities

IRI

Total Data Management

CELEBRATING 40 YEARS 2018

The CoSort Company